Lab 8: midterm review

Statistics for Social Scientists II

Bur, GJM

2024-10-28

1 Algebra and conceptual practice

1. Write the formula for the general sample variance and explain it in intuitive terms. Then, consider what you would do if I asked you for the "conditional variance" of Y in the sample, given that $X = 4$. Finally, keeping in mind that a linear regression assumpton is that the conditional mean of the population is given by a linear function of the predictors, write out the conditional variance formula and explain how it is related to the simple idea of conditional variance given above. Explain also how this conditional variance is related to one of the sums of squares in our "generalized Pythagorean theorem" ($SS_T = SS_M + SS_E$).
2. True or false? If a 95 percent confidence interval about a population mean (let's say) μ_Y given a sample mean $\bar{y}$ includes some particular value k, then we would fail to reject $H_0 : \mu_Y = k$ using a *one-tailed test*.
3. From the following summary statistics and variance/covariance matrix from Galton's famous height data, calculate the bivariate regression slope and intercept for the following data. Then, find the predicted height of a child whose parents' average height was 60 inches and then 70 inches.

	midparent	height
mean	66.64	66.74
std	1.74	3.58

	midparent	height
midparent	3.02	2.01
height	2.01	12.82

4. Calculate the fully standardized regression slope and compare it to the correlation coefficient.
5. The model sum of squares is 1242.83; the in-sample error sum of squares is 10708.74. Find the model F-statistic, and explain what it means generally and in this case. Since

in the bivariate case $F = t^2$, what can we say about whether or not these variables are related in the population?

6. Does the intercept matter in this model? If not, why include it anyways?

7. Suppose we now include the gender of the child as a predictor in the model and center the "midparent height" at the mean. We obtain the following results. Interpret the coefficients and the confidence intervals. Compare R^2 from this model to the previous model; which is better?

	0	1	2	3
0	Model:	OLS	Adj. R-squared:	0.634
1	Dependent Variable:	height	AIC:	4093.76
2	Date:	2024-10-26 20:23	BIC:	4108.28
3	No. Observations:	933	Log-Likelihood:	-2043.9
4	Df Model:	2	F-statistic:	807.6
5	Df Residuals:	930	Prob (F-statistic):	4.83e-204
6	R-squared:	0.635	Scale:	4.6957

	Coef.	Std.Err.	t	P>	t	
const	64.058	0.101845	628.975	0	63.8581	64.2578
midparent_c	0.717334	0.0408824	17.5463	9.30625e-60	0.637102	0.797567
male	5.21965	0.142033	36.7496	2.44868e-183	4.94091	5.49839

8. Now, let's see if birth order matters. We could treat birth order as a continuous variable, but it might make more sense to turn it into a categorical variable. This is the late 19th century, so people had lots of kids; let's say we want to see if being born first, being born second, or being born any later matters. Write out our null hypothesis for this test of a *set* of predictors (first born and second born) clearly.

	0	1	2	3
0	Model:	OLS	Adj. R-squared:	0.676
1	Dependent Variable:	height	AIC:	3980.49
2	Date:	2024-10-26 20:29	BIC:	4004.68
3	No. Observations:	933	Log-Likelihood:	-1985.2
4	Df Model:	4	F-statistic:	488
5	Df Residuals:	928	Prob (F-statistic):	1.96e-226
6	R-squared:	0.678	Scale:	4.15

	Coef.	Std.Err.	t	P>	t	
const	63.872	0.0981767	650.582	0	63.6793	64.0646
midparent_c	0.703069	0.038469	18.2762	5.66011e-64	0.627573	0.778565

	Coef.	Std.Err.	t	P>	t	
male	4.37369	0.15823	27.6413	3.54694e-123	4.06316	4.68422
firstborn	2.11799	0.189991	11.1479	3.57158e-27	1.74513	2.49085
secondborn	0.845285	0.19475	4.34035	1.57877e-05	0.463083	1.22749

Calculate the partial F-statistic for the inclusion of the birth-order dummies using the following information. Recall that you can do this simply by taking the increase in the model sum of squares between models over the increase in the number of predictors (the mean model square for the incremental predictors), all divided by the error sum of squares for the full divided by its df (the mean error square for the full model). Interpret the results.

	model2	model3
SS_M	7584.55	8100.35
SS_E	4367.02	3851.23
SS_T	11951.6	11951.6

9. Find the predicted height for a first-born male whose parents were $5'5''$ and $5'11''$. Find the predicted height of a second-born female whose parents were $6'1''$ and $5'2''$. These are predicted heights for me and my fiancée; now, find our residuals if my actual height is $6'2''$ and hers is $5'9''$. Be careful! Consider what our intercept represents here.

2 Stata practice

10. Replicate the first regression above *by hand* using the data. To import the data, which are in comma-separated values (CSV) format, write `import delimited using $filepath to data$`. To make a birth-number variable, assuming that the data are already sorted by birth order within families, use the following: `bysort family: egen birthnumber = rank(_n)`. You might find it useful to avail yourselves of the fact that …

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

11. Replicate the standard error of the regression slope by hand. Use the fact that…

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

…where in this *specific* instance $\hat{\sigma}^2$ is the conditional variance about the regression line.

12. Replicate the results for Exercise 7 above; you may now use `reg`. You can either `destring gender` or just use a string expression, e.g. `gen male if gender == "male"`.

13. Replicate the results from Exercise 8 above. Consider carefully how to make a set of dummies encoding birth order and check your work. (At the end you might try running `table birthnumber firstborn secondborn`). Replicate the partial F-test as well.

14. Replicate the results from Exercise 9 above. Try using `margins` if possible, but failing that, use `matrix list e(b)` after `reg` and access elements of `e(b)` with `e(b)[1,1]` (for example).

15. Is it 2016 all over again? Download the American National Election Survey extract from the course Github. Make a scale indicating someone's approval of public spending. Do this in the way Prof. Gerber recommends—data are precious, so keep anyone who responded to any item. Consider dropping some items. Do we need to standardize here? Why or why not? If we do standardize, does the formula become ever more nicely comprehensible as a function of test length k and average inter-iterm corelations $\hat{\mu_\rho}$? Should we drop some items? Finally, regress someone's approval of the Democratic Party on the scale you create and note anything interesting.